ÉCOLE NATIONALE DES CHARTES

Paul, Hector Kervegan

Traitement, exploitation et analyse d'un corpus semi-structuré : le cas des catalogues de vente de manuscrits

Mémoire pour le diplôme de master

Technologies numériques appliquées à l'histoire

2022

Résumé

Introduction

J'ai choisi de structurer mon mémoire autour de plusieurs questions connexes, qui, à différents degrés, se retrouvent tout au long du développement :

En quoi la nature semi-structurée du corpus permet d'en automatiser le traitement ? Comment produire des informations normalisées et exploitables à partir d'un corpus textuel semi-structuré ? En quoi ce traitement et la traduction des documents vers d'autres formats et d'autres médias impacte leur réception ? Quels sont les choix techniques qui influencent cette réception ?

Les deux premières questions, d'orientation plutôt technique, forment la colonne vertébrale pour le mémoire ; elles lient deux aspects centraux : la nature du corpus et la manière dont sa structure permet toute la chaîne de traitement. Par « semi-structuré », j'entends que, à un niveau distant, toutes les entrées de catalogue suivent la même structure ; des séparateurs distinguent les différentes parties, et les informations sont souvent structurées de manière semblable pour chaque manuscrit vendu. Cela permet un traitement de « basse technologie » (*low-tech*) en évitant d'entraîner de lourds modèles de traitement du langage naturel (ce qui aboutirait à des solutions complexes, difficiles à maintenir et à faire évoluer et relativement opaques dans leur fonctionnement). À l'inverse, un corpus semi-structuré peut être traité en déduisant une « structure abstraite », que chaque entrée de catalogue partage. Il est alors possible de mettre en place des solutions techniques plus faciles, pour un résultat de qualité équivalente. Produire des « informations normalisées et exploitables » implique de traiter le corpus en cherchant des réponses à des questions de recherche précises – dans le cadre de mon stage, une question centrale a été de chercher à isoler les facteurs déterminant le prix d'un manuscrit.

Les deux dernières questions, au premier abord plus théoriques, me semblent centrales, notamment à la troisième partie de ce mémoire. Numérisation, traitement informatisé et diffusion sur le web ne sont pas des opérations neutres, mais un ensemble de « traductions » des documents originels. Ces processus comportent une part de choix conscients, qu'il s'agit de mettre en avant. Par exemple, on considère que la majorité des documents vendus ont pour titre l'auteur.ice du document. Cette personne n'est cependant pas toujours mentionnée, et des documents peuvent être nommés d'après un lieu, un évènement ou un thème (la révolution française, par exemple). Ces « traductions » des catalogues sont relativement discrètes tout au long de la chaîne de traitement (où le

format dominant est la `TEI`, qui garde une relation d'équivalence avec le texte). C'est lors du passage au site web que ce processus de traduction devient plus évident, et, potentiellement, plus problématique. On y abandonne la référence au document originel (les catalogues numérisés ne sont pas accessibles en ligne par un.e utilisateur.ice), le catalogue n'est plus la manière privilégiée d'accéder aux items vendus... De plus, la construction d'un site web implique la conception d'une interface et, dans notre cas, la production d'une série de visualisations intégrées au site. Le passage au site web remet aussi en cause la hiérarchie habituelle entre ingénierie et recherche : la conception d'un site ne répond pas à une question scientifique, mais elle soulève ses propres questions. Loin d'être anodines, ces problématiques de design déterminent la construction et la réception des savoirs. Il est donc important, je pense, de problématiser ces questions de visualisation et de design.

Première partie

Du document numérisé au XML-TEI : nature du corpus, structure des documents et méthode de production des données

Chapitre 1

Présentation du corpus

Ce chapitre est dédié à une présentation des documents traités dans le cadre du projet MSS : nature, quantité de documents (et d'entrées individuelles), dates et différents types de catalogues vendus. On pourra également représenter la répartition des ventes par an grâce aux graphiques produits pour le site web avec Plotly (à moins que cela ait plus de sens en troisième partie). Ce chapitre s'appuie sur les mémoires effectués par d'ancien.ne.s stagiaires de Katabase, qui ont déjà beaucoup analysé la nature et les enjeux du corpus [1].

1.1 Intérêt scientifique des catalogues de vente

1.2 La structure du corpus : périodisation, producteurs des documents et classification

aka quand et en quelle masse les documents sont produits, qui sont les producteur.ices (Charavay, RDA... ATTENTION À PAS FAIRE DE CONFUSION AVEC LE TERME DE PRODUCTEUR ON EST RELU PAR DES ARCHIVISTES) et quelle catégorisation entre les catalogues on peut faire (LAV, RDA / prix fixes et prix non fixes...)

1. Lucie Rondeau du Noyer, *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Charavay*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : `https://github.com/lairaines/M2TNAH` (visité le 13/06/2022) ; Caroline Corbières, *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH` (visité le 13/06/2022) ; Juliette Janès, *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

1.3 La structure des catalogues

(parler des catalogues au niveau du catalogue complet, pas de la page ; voir si on parlerait pas plutôt de ça dans la partie suivante)

Chapitre 2

Production des données : océrisation et structure des documents traités

Cette partie s'attache autant à présenter le processus d'océrisation (qui est déjà bien établi et ne constitue pas le cœur de mon stage) que la structure des documents. Alors que le chapitre précédent s'intéresse aux catalogues dans leur ensemble, ici, on étudie le corpus au niveau de la page et de l'entrée individuelle. En effet, l'océrisation repose sur la segmentation, et donc sur l'établissement d'une structure « abstraite » d'une page (c'est-à-dire, d'un découpage de la page en zones).

Chapitre 3

Du texte à la TEI : méthode de transformation des documents océrisés en fichiers XML-TEI valides

Après une étape d'océrisation via eScriptorium, le texte extrait des PDF peut être exporté soit en texte brut, soit en XML Page ou Alto. Ces formats s'attachent à garder une relation entre le XML et le document numérisé (les zones de texte sont indiquées, chaque ligne est dans une balise...). Cependant, l'unité intellectuelle centrale à la suite du projet, ce n'est pas la page numérisée, mais l'entrée de catalogue. Un format plus complexe que le XML d'eScriptorium est donc nécessaire. Assez logiquement, la suite du projet s'appuie sur une traduction des catalogues en TEI.

Jusqu'à maintenant, cette transformation était faite par GROBID-Dictionnaries et des feuilles XSL. Le projet GROBID-Dictionnaries étant maintenu de façon assez opaque, le choix a été fait de remplacer cet outil par une solution plus simple (qui ne passe pas par le *machine learning*) et modulable. En s'appuyant sur la nature « semi-structurée » du corpus, il est possible de séparer les différentes parties du texte en identifiant des séparateurs. Ce chapitre s'intéresse donc à la transformation de Alto vers la TEIà l'aide d'*expressions régulières*. En changeant d'unité intellectuelle – ou en choisissant de privilégier une unité au sein des catalogues plutôt qu'une autre (page, série d'entrées...) –, on prend de la distance et on interprète le catalogue papier.

3.1 Points de départ pour une transformation en TEI

3.1.1 L'Alto, base du processus de « traduction »

Ici, on présente la structure du Alto produit par eScriptorium, qui est la base à partir de laquelle la transformation en TEI est faite.

3.1.2 Au delà des manuscrits : dériver une structure pour les entrées afin de permettre la transformation des fichiers

Cette sous-section s'intéresse à la structure des entrées individuelles, à deux niveaux :

— Au niveau intellectuel : quelles sont les différentes parties d'une entrée (titre, description du manuscrit, prix...).

— Au niveau « textuel » : quels sont les séparateurs, c'est à dire les éléments dans le texte qui permettent de séparer les pages de catalogue en entrées et les entrées en sous-éléments) correspondant à la structure intellectuelle décrite ci-dessus.

3.2 Le résultat à produire : présentation de l'édition TEI des catalogues de vente de manuscrits

Ici, on présente le résultat attendu de la transformation Alto, soit une édition numérique en TEI des catalogues de vente. On s'intéresse autant à la structure des documents XML (quelles balises sont utilisées...) qu'à l'intérêt scientifique d'une édition numérique (balisage sémantique, possibilité de normaliser les informations grâce à des attributs).

3.3 Processus de « traduction » des documents de l'Alto à la TEI

Cette section s'intéresse au processus technique de transformation des documents de l'Alto vers la TEI.

Deuxième partie

Normalisation, enrichissements et extraction d'informations : une chaîne de traitement pour des données semi-structurées

Chapitre 4

Homogénéiser et normaliser un corpus complexe

Ce chapitre s'intéresse à la manière dont les fichiers `TEI` sont traités afin de pouvoir ensuite en extraire des informations. C'est directement grâce la structure des entrées (et grâce à la nature « semi-structurée » des catalogues) qu'est possible le traitement automatisé des documents. Cette partie s'attache également à rappeler les questions de recherche qui sous-tendent la normalisation des documents (ajouter plus de structure au document `TEI` pour l'exploiter plus facilement, uniformiser la notation des tailles et des dimensions des documents...). Cette partie sera probablement relativement brève, puisque l'étape de normalisation des données a déjà été faite par d'autres stagiaires et documentée d'autres mémoires.

Chapitre 5

Extraction d'informations

Ici sont décrits le processus et les objectifs de l'extraction d'informations à partir des fichiers `TEI`. Des données sont extraites pour chaque entrée de catalogue (un travail largement effectué par A. Bartz, que j'ai légèrement mis à jour) : prix (dans la monnaie de l'époque et en francs constants), date de vente normalisée, nom de l'auteur.ice et description du manuscrit... Un second processus d'extraction produit des données pour chaque catalogue de vente : titre du catalogue, date de vente, nombre d'items vendus, prix minimum, inférieur et maximum, prix moyen et médian, variance... En même temps que ces processus d'extraction d'informations, un script de conversion des monnaies (françaises et étrangères) en francs constants 1900 a été élaboré. En annulant l'effet de l'inflation, les francs constant permettent d'étudier l'évolution réelle des prix.

L'extraction d'informations pour les entrées individuelles permet surtout de faire une réconciliation des manuscrits vendus (c'est à dire, de retrouver les items vendus plusieurs fois). Le deuxième processus permet de produire des données statistiques sur l'évolution du cours et du volume du marché des manuscrits (nombre d'items en vente, évolution des prix) ; c'est à partir de ces informations normalisées que sont construites les visualisations intégrées au site de Katabase.

Chapitre 6

Vers une étude des facteurs déterminant le prix des documents : alignement des entrées du catalogue avec Wikidata et exploitation de données normalisées

Ce chapitre est construit autour d'une question de recherche : comment produire des informations exploitables pour une étude économétrique à partir d'un corpus textuel semi-structuré ? Un des objectifs du projet est de faire l'étude des facteurs déterminant le prix d'un manuscrit. Pour faire cette étude, il faut obtenir, pour chaque entrée du catalogue, un certain nombre d'informations normalisées. Le travail d'extraction de données présentes dans les catalogues a déjà été fait par de précédent.e.s stagiaires. Ces données sont principalement quantitatives : prix des manuscrits, dimensions et nombre de pages, date de création. Il est nécessaire de compléter les informations par des données qualitatives et d'enrichir les données disponibles avec des sources extérieures. Pour ce faire, il a été choisi d'aligner le nom des auteur.ice.s des manuscrits avec des identifiants Wikidata ; dès lors que l'on a un identifiant Wikidata, il est possible de récupérer automatiquement des informations sur les personnes via `SPARQL`. Le choix de travailler uniquement sur les noms, et non sur la description des documents, a deux motivations :

— Les noms de personnes (et la manière dont elles sont décrites) constituent la partie la plus normalisée des documents. La description des manuscrits est plutôt en « texte libre ». Dans la continuité avec le reste du projet, nous sommes resté dans une approche « basse technologie », qui consiste à s'appuyer majoritairement sur des solutions techniquement simples. C'est pourquoi nous avons préférer traiter

les noms avec des tables de correspondance [1] et des *expressions régulières*, plutôt que de faire du TAL sur la description des documents.

— Toutes les informations « simples » (données quantitatives facilement normalisables : dates etc.) ont déjà été extraites des descriptions des manuscrits.

Ce travail d'enrichissement a été fait en deux temps.

La première étape, et la plus difficile, est l'alignement avec Wikidata. Cela demande d'extraire un ensemble d'informations à partir du nom de la personne et de la description de celle-ci (nom, prénom, titre de noblesse, occupation, dates de vie et de mort...). À partir de ces informations, stockées dans un dictionnaire, un algorithme construit successivement différentes chaînes de caractères à rechercher sur l'API de Wikidata. L'objectif est que le premier résultat recherché sur Wikidata soit correct. Sur un jeu de test, le score F1 [2] obtenu est de 68%. Un relecture « manuelle » des résultats est donc nécessaire.

La deuxième étape, nettement plus simple, consiste à lancer des requêtes Wikidata sur les identifiants récupérés afin de récupérer des informations sur les auteur.ice.s des manuscrits (cette partie du travail est encore en cours) pour enrichir nos données.

Une fois ce travail effectué, l'enrichissement des données à proprement parler est possible : les fichiers `TEI` sont mis à jour pour ajouter les identifiants Wikidata. Ainsi, il est possible de faire le lien entre les entrées de catalogues dans des fichiers `XML` et les données issues de requêtes `SPARQL`, stockées dans un `JSON`.

6.1 Questions introductives : pourquoi et comment s'aligner avec Wikidata ?

Cette section, introductive, répond à une question évidentes mais essentielle, puisqu'elles permettent de mettre au clair l'intérêt et les (multiples) difficultés dans l'alignement avec Wikidata.

6.1.1 Pourquoi s'aligner avec des identifiants Wikidata ?

Nos données sont déjà complètes, une pipeline entière existe déjà. Cependant, il peut être difficile de déterminer ce qui fait le prix d'un manuscrit. On aborde les manuscrits avec nos propres catégories intellectuelles du XXIème siècle, et notre connaissance de l'histoire de l'époque. Il n'est pas non plus possible de reconstruire d'une manière exacte

1. C'est à dire, des tables qui permettent de normaliser la manière dont les informations figurent dans les catalogues, et donc de remplacer des termes « vernaculaires » par leurs équivalents utilisés par Wikidata

2. Le score F1, ou *F-score*, est la moyenne harmonique de la précision (vrais positifs par rapport au total de résultats obtenus) et du rappel (nombre de résultats positifs par rapport au total de résultats positifs). (*Précision et rappel*, Wikipedia. L'encyclopédie libre, 2022, URL : `https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel` [visité le 13/06/2022])

le regard qu'un public du XIXème siècle aurait sur ces manuscrits – ce qui permettait de revenir à une perception antérieure de la valeur. Il faut donc chercher à contourner ces biais en produisant des données aussi objectives que possibles. Ainsi, un maximum de variales sont à notre disposition pour calculer des régressions linéaires (qui permettent de prédire l'impact d'une variable sur l'évolution des prix).

6.1.2 Comment traduire des descriptions textuelles datant du XIXème siècle en chaînes de caractères qui puissent retourner un résultat sur Wikidata ?

Ce problème est autant linguistique de technique. Une personne ou une chose est nommée ou décrite d'une certaine manière dans un catalogue de vente ancien. Il n'y a aucune garantie que cette caractérisation corresponde à ce qui est disponible sur Wikidata : l'orthographe des noms évoluent, tout comme la manière de nommer certains métiers. À ces évolutions graphiques s'ajoutent des évolutions intellectuelles : les titres de noblesse sont un marqueur plus important au XIXème siècle français que dans un XXIème siècle mondialisé. Une personne n'est que rarement décrite par son titre dans Wikidata.

6.1.3 Comment négocier avec le moteur de recherche de Wikidata ?

Si les catalogues de vente fonctionnent avec leurs propres catégories mentales, le même peut être dit de Wikidata : certains types de données sont plus souvent référencées que d'autres et Wikidata utilise un vocabulaire qui lui est propre. Par expérience, le moteur de recherche de Wikidata est assez « rigide » : contrairement à un moteur généraliste, il n'admet pas d'orthographes alternatives, par exemple. Le traitement des données textuelles et tout le processus de normalisation des données dépendent de ces faits : il faut trouver quelles informations sont référencées par Wikidata et comment elles sont référencées, et développer une méthodologie pour obtenir les meilleurs résultats possibles.

6.1.4 Quelles données extraire via `SPARQL` ?

Dans cette sous-section, nous détaillons les données récupérées via `SPARQL` ainsi que les choix scientifiques qui sous-tendent nos décisions.

6.2 Préparer et structurer les données

Avant de chercher à récupérer un identifiant Wikidata via l'API, un algorithme se charge de traduire et de structurer les données : à partir d'un nom et de son éventuelle des-

cription, un dictionnaire qui contient les informations de manière structurée est construit. À partir de ce dictionnaire, un algorithme contenant différentes requêtes est lancé pour récupérer les identifiants Wikidata.

6.2.1 Présentation générale

Ici, on présente la pipeline de l'algorithme (à l'aide d'un schéma), les données fournies en entrée et le résultat produit en sortie. Les sections suivantes détaillent quelques points d'intérêts.

6.2.2 Identifier le type de nom

Les éléments `tei:name` contiennent le nom qui est donné à un document. Si c'est souvent un nom de personne, ce n'est pas toujours le cas (il y a aussi des noms de lieux, d'évènements), et il y a plusieurs types de noms de personnes : un nom peut être écrit en suivant différentes structures, ce qui appelle à différents types de traitements.

6.2.3 Reconstruire un prénom complet à partir de son abréviation

Souvent, le prénom d'une personne est écrit en abrégé. Partant de ce constat, un algorithme a été construit pour :

— Repérer lorsqu'un prénom est abrégé, en prenant en compte différents types d'abréviations (nom simple ou composé, nom entièrement ou partiellement abrégé) et des possibles fautes dans les catalogues (un point est oublié à la fin d'une abréviation, par exemple).

— Reconstruire un prénom complet à partir de son abréviation, ce qui passe pas un algorithme qui cherche à reconstruire le nom en plusieurs étapes pour obtenir le nom le plus complet possible avec un mimimum d'erreurs.

6.2.4 Extraire des informations normalisées à partir d'un nom et de sa description

Cette sous-section détaille l'utilisation de tables de conversion pour traduire et normaliser certaines donnes importantes (dates de vie et mort, titres de noblesse et fonctions).

6.3 Extraire des identifiants Wikidata

Une fois un dictionnaire de données normalisées produites, un algorithme lance des recherches en plein texte sur l'API de Wikidata afin de récupérer des identifiants. L'algo-

rithme lance plusieurs requêtes successivement. L'objectif est de récupérer un identifiant en lançant le moins de requêtes, avec le plus de certitude possible (ce qui implique de quantifier la certitude).

6.3.1 Présentation générale

Ici est présenté le fonctionnement général de l'algorithme, qui se comporte différemment en fonction du type de données qu'il a à traiter (personne noble ou non, prénom reconstruit ou non...)

6.3.2 Gérer la montée en charge : optimisation et réduction du temps d'exécution

Le script est assez compliqué, repose sur une API et traite un grand nombre de données (plus de 82000 entrées). Il prend donc plus d'une dizaine d'heures à s'exécuter et demande des performances matérielles élevées (la première version du script ne fonctionnait plus sur mon ordinateur après avoir traité 5% du jeu de données). Son optimisation nécessaire est donc décrite dans cette sous-section

6.3.3 Évaluation du script : tests, performance et qualité des données extraites de Wikidata

Des tests ont été réalisés pour :

— isoler l'impact de chaque paramètre (élément du dictionnaire) dans l'obtention des bons résultats
— évaluer la qualité de l'algorithme final
— mesurer la performance de celui-ci.

Ces tests, et leurs résultats, sont présentés ici.

6.4 Après l'alignement, l'enrichissement : utiliser SPARQL pour produire des données structurées

6.4.1 Produire des données exploitables via SPARQL

La récupération des identifiants Wikidata est la partie la plus complexe dans l'utilisation de Wikidata pour enrichir des données. Après un rappel sur les informations extraites, le processus d'extraction d'informations et de stockage dans un JSON est détaillé.

6.4.2 Rendre les données exploitables

Lier la `TEI` aux données nouvellement produites

Cette courte section détaille la mise à jour des fichiers `TEI` avec les identifiants Wikidata, ce qui permet de faire le lien entre les entrées de catalogues et les données issues de Wikidata.

Produire des données permettant d'étudier les facteurs déterminant la valeur d'un manuscrit

Ici est détaillé le document produit pour calculer des régressions linéaires sur les manuscrits, et chercher à identifier les valeurs déterminant l'évolution des prix.

6.5 Des données à la monnaie : premiers résultats de l'étude

Sous réserve que l'étude des régressions linéaires ait été fait à temps (ce qui n'est pas garanti), j'aimerais ici présentés les premiers résultats sur les facteurs de l'évolution des prix.

Troisième partie

Après la TEI : l'application web *Katabase*, interface de diffusion des données

Chapitre 7

Design d'interface dans un projet d'humanités numériques : l'application web *Katabase*

Ce chapitre s'intéresse aux relations entre *web design*, données textuelles et humanités numériques, à partir de l'exemple du site web développé pour le projet *Katabase*.

7.1 Le design d'interfaces : une reconfiguration des méthodes de recherche et une transformation du corpus

Cette section s'intéresse aux nouveautés apportées par le design d'interfaces dans les humanités numériques. On s'intéresse à la manière dont le design d'interfaces (et le design de façon générale) transforme les méthodes de recherche « habituelles », mais aussi une transformation du rapport aux documents.

7.1.1 Le design comme inversion des méthodes

Avec les humanités numériques, les questions de design et de structuration deviennent centrales, depuis la conception de schémas `TEI` (qui demandent de mettre en forme un document pré-existant) et d'ontologies jusqu'au développement d'interfaces et de sites web. Parmi ces questions « formelles », le design d'interfaces occupe cependant une place particulière. En effet, dans la plupart des aspects des humanités numériques, le rapport entre questions techniques et scientifiques est clairement établi ; la question scientifique préexiste, et la technique sert surtout à répondre à cette question (comme cela a été le cas jusqu'à dans la « pipeline » jusqu'ici). Cette hiérarchie entre théorie et

pratique reste somme toute assez traditionnelle et correspond aux méthodes scientifiques établies.

Avec le design d'interfaces, cependant, ce rapport établi s'inverse. En effet, le design ne cherche pas à répondre à une question. Tout au plus, il répond à un cahier des charges (le design doit, à minima, permettre de diffuser des données de façon lisible par des êtres humains). C'est avec la pratique du design que naissent les problématiques, parmi lesquelles :

— Comment organiser les différentes parties d'une page pour que celle ci soit lisible ?
— Comment organiser la relation entre les pages pour qu'un site web soit facilement navigable ?
— De quelle manière l'apparence d'un site détermine la réception des contenus ?
— En quoi le design d'un site web construit ou bouscule des habitudes et des formes d'utilisation chez ses utilisateur.ice ?

Toutes les questions posées par le design n'attendent pas nécessairement de réponse. Cependant, force est de constater que ce domaine appelle à une nouvelle approche pour des chercheur.euse.s et ingénieur.e.s issu.e.s des humanités ; ces questions visuelles amènent à une approche semblable à celle de la recherche-création et demandent de développer un nouveau rapport à la technique.

7.1.2 Interface et document

En plus de perturber nos méthodes, la conception d'interfaces influence la perception des documents. Dans le cas du projet *Katabase*, le site web opère une médiation, il implique de une « scénographie » autour des catalogues de vente. Ceux-ci et les manuscrits qui y ont décrits sont intégrés à des pages, inclus dans un parcours, accessibles depuis différents points d'entrée. En plus de cette scénographie, les catalogues sont littéralement traduits, depuis la `TEI` vers le format `HTML`, ce qui implique une perte d'information (les métadonnées du `teiHeader`). Enfin, le site internet marque avant tout un éloignement intellectuel avec les documents : le catalogue n'y est plus l'unité intellectuelle dominante, alors qu'il restait l'un des critères structurants des fichiers `TEI` (un fichier représentant un catalogue). Sur le site web, on peut accéder directement aux éléments vendus, sans avoir à passer par les catalogues. Dans le contexte d'un projet issu de la littérature, toutes ces opérations ne sont pas neutres et méritent d'être explicitées. La méthode de traduction de `XML-TEI` vers `HTML` peut également être présentée ici.

7.2 La conception d'interface, un problème pour les humanités numériques ?

Cette section s'intéresse aux rôle des interfaces en humanités numériques.

7.2.1 L'interface comme méthode de communication

7.2.2 Pour une approche pragmatique du design d'interfaces dans un contexte d'humanités numériques

Le design graphique demande des compétences spécifiques qui ne font pas directement partie des cursus d'humanités numériques. Cependant, le design ne sert pas seulement à faire des sites qui soient « beaux », il joue un rôle essentiel en encadrant la réception des contenus présentés. Cependant, les approches plus « élaborées » de design d'interfaces demandent des financements et des techniques qui sont souvent hors de portée d'un projet universitaire. Des approches plus « critiques » du design ont également été développées dans les humanités numériques [1]. Ces approches ont tendance à être difficiles à mettre en œuvre ; leur portée critique peut aller à l'encontre de l'utilité des interfaces, en faisant de l'interface l'objet principal d'intérêt, aux dépends des contenus présentés.

À l'opposé de ces approches, ce qui est défendu dans le cadre du projet *MSS / Katabase* est une approche à la fois informée et pragmatique du *web design*. Informée, car être conscient des enjeux du design permet un meilleur positionnement en tant qu'ingénieur.e, et donc une présentation des contenus plus intéressante. Pragmatique, parce que les solutions qui sont présentées sont des solutions techniquement réalisables dans le cadre d'un projet universitaire. C'est ici qu'est présentée la charte graphique développée pour l'application web *Katabase*.

7.2.3 Rejeter les interfaces ?

Après avoir parlé de l'intérêt des interfaces et présenté l'approche suivie au sein du projet *MSS / Katabase*, cette partie s'attache à développer une critique des interfaces. À partir d'une approche historique des interfaces graphiques, des contextes dans lesquelles elles se sont développées, nous revenons sur les concepts centraux à leur développement que sont la notion d'utilisateur et de design d'expérience. Il ne s'agit pas de remettre en cause l'utilisation d'interfaces, mais de défendre une approche critique et consciente de l'impact que la standardisation des « expériences utilisateur » sur internet peuvent avoir sur la diffusion des connaissances.

1. Johanna Drucker, *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

Chapitre 8

Visualisations

Dans cette partie, je m'intéresserai plus spécifiquement aux problématiques techniques et scientifiques liées à la visualisation de données, qui ont constitué une des missions centrales de mon stage :

— Qu'est-ce que l'on cherche à montrer avec ces visualisations ?

— Comment intégrer ces visualisations de manière « élégante » et efficace au site ? Par exemple, 7 types de graphiques sont produits pour l'index des catalogues. Il faut donc trouver un moyen de tous les présenter, sans pour autant alourdir excessivement la page. La visualisation doit également être pensée en concertation avec le reste du design du site web (largement mis à jour pendant mon stage), et s'intégrer à sa charte graphique.

— Comment construire ces visualisations ? Comment faciliter la lecture d'informations complexes ? Quel impact les visualisations ont-elles sur la perception des informations ? Pour ces problématiques techniques, je pense (entres autres) m'appuyer sur le mémoire de Ségolène Albouy [1].

1. Ségolène Albouy, *Médiation des données de la recherche. Élaboration d'une plateforme en ligne pour une base de tables astronomiques anciennes.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : `https://github.com/Segolene-Albouy/Memoire-TNAH2019` (visité le 13/06/2022).

Chapitre 9

Visualisation et interfaces : un problème pour les humanités numériques ?

Ce chapitre de conclusion, plus théorique, revient sur les débats actuels concernant le lien entre interfaces graphiques, visualisation de données et humanités numériques. En s'appuyant (entre autres) sur les théories de Johanna Drucker [1] et Anthony Masure [2], il s'attache à réintégrer les questions « visuelles » propres aux humanités numériques à un contexte plus large. Le chapitre revient sur les origines des interfaces graphiques (et donc sur les objectifs implicites qui structurent nos interactions avec les machines) ; il cherche à remettre la visualisation dans les humanités numériques en lien avec une tendance globale à la visualisation – tendance qui vient du monde de l'entreprise, ce qui n'est pas anodin. Du fait de la quantité de significations implicites sous-jacentes à la construction d'interfaces, il est, je pense, nécessaire d'avoir une approche théorique et critique du rapport des humanités numériques aux problématiques de design et aux « dispositifs » [3] que sont nos outils de travail.

- interface web : un objet problématique et une nouvelle interprétation des documents (en gros, sur l'historique des interfaces + sur ce qui arrive aux documents quand on les transforme en un site web)

- le design : un nouveau problème pour les humanités ? en gros, je dis que le web design et la question d'esthétique est une question nv pour les humanités "non-créatrices", où la forme reste assez secondaire à l'idée ; ici, on est dans un entre-deux, où penser la forme est nécessaire. remettre ça en question de 2 manières : - une histoire de la visualisation de données et du design en sciences avec A.L. Renon ; - le fait que s'intéresser à des questions

1. J. Drucker, *Visualisation : l'interprétation modélisante…*
2. Anthony Masure, *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01).
3. Giorgio Agamben, *What is an apparatus ? and Other Essays*, Première édition, Stanford, 2009 (Meridian : Crossing aesthetics).

de design dans les DH, sans vraiment d'approche critique ou de connaissance des enjeux du design, ça vient avec 2 dangers : - faire des interfaces qui desservent le contenu, en les traduisant mal ou en figeant les humanités dans un langage esthétique dépassé - faire du web design avec une position acritique, c'est risquer de suivre les dominantes, et donc de copier un langage esthétique qui est déterminé par le système économique - une partie sur la visualisation comme interprétation - pour une approche critique du design, mon ptit plaidoyer final

Bibliographie

À propos du projet Katabase / MSS

Corbières (Caroline), *Du catalogue au fichier TEI. Création d'un workflow pour encoder automatiquement en XML-TEI des catalogues d'exposition.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2020, URL : `https://github.com/carolinecorbieres/Memoire_TNAH` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie) et Khemakhem (Mohamed), « Selling autograph manuscripts in 19th c. Paris : digitising the Revue des Autographes », dans *IX Convegno AIUCD*, Milan, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02388407` (visité le 13/06/2022).

Gabay (Simon), Rondeau du Noyer (Lucie), Gille Levenson (Matthias), Petkovic (Ljudmila) et Bartz (Alexandre), « Quantifying the Unknown : How many manuscripts of the marquise de Sévigné still exist ? », dans *Digital Humanities DH2020*, Ottawa, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02898929/` (visité le 13/06/2022).

Janès (Juliette), *Du catalogue papier au numérique. Une chaîne de traitement ouverte pour l'extraction d'information issue de documents structurés*, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2021, URL : `https://github.com/Juliettejns/Memoire_TNAH` (visité le 13/06/2022).

Khemakhem (Mohamed), Romary (Laurent), Gabay (Simon), Bohbot (Hervé), Frontini (Francesca) et Luxardo (Giancarlo), « Automatically Encoding Encyclopedic-like Resources in TEI », dans *The annual TEI Conference and Members Meeting*, Tokyo, 2018, URL : `https://hal.archives-ouvertes.fr/hal-01819505` (visité le 13/06/2022).

— « Information Extraction Workflow for Digitised Entry-based Documents. » Dans *DARIAH Annual event 2020*, Zagreb, 2020, URL : `https://hal.archives-ouvertes.fr/hal-02508549` (visité le 13/06/1997).

Rondeau du Noyer (Lucie), *Encoder automatiquement des catalogues en XML-TEI. Principes, évaluation et application à la revue des autographes de la librairie Cha-*

ravay, Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/lairaines/M2TNAH (visité le 13/06/2022).

Rondeau du Noyer (Lucie), Gabay (Simon), Khemakhem (Mohamed) et Romary (Laurent), « Scaling up Automatic Structuring of Manuscript Sales Catalogues », dans *TEI 2019 : What is text, really ? TEI and beyond*, Graz, 2019, URL : https://hal.archives-ouvertes.fr/hal-02272962 (visité le 13/06/2022).

Visualisation et design d'interfaces

Agamben (Giorgio), *What is an apparatus ? and Other Essays*, Première édition, Stanford, 2009 (Meridian : Crossing aesthetics).

Albouy (Ségolène), *Médiation des données de la recherche. Élaboration d'une plateforme en ligne pour une base de tables astronomiques anciennes.* Mémoire pour le diplôme de master "Technologies numériques appliquées à l'histoire", Paris, École nationale des Chartes, 2019, URL : https://github.com/Segolene-Albouy/Memoire-TNAH2019 (visité le 13/06/2022).

Drucker (Johanna), *Visualisation : l'interprétation modélisante*, Paris, 2020 (Esthétique des données, 03).

Masure (Anthony), *Design et humanités numériques*, Paris, 2017 (Esthétique des données, 01).

Économétrie et statistiques

Piketty (Thomas), *Les hauts revenus en France au XXe siècle : inégalités et redistributions, 1901-1998*, Paris, 2001.

Précision et rappel, Wikipedia. L'encyclopédie libre, 2022, URL : https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel (visité le 13/06/2022).

Table des matières